1. In the sense of machine learning, what is a model? What is the best way to train a model?

🡪A Machine Learning Model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Model Naming — Give Your Model a Name, Let’s start with giving your model a name, describe your model and attach tags to your model. Tags are to make your model searchable.

Data Type Selection — Choose data type(Images/Text/CSV), It’s time to tell us about the type of data you want to train your model. ML Models support Images, Text and \*.CSV (categorical data) data types.

Data Upload — Upload your data or choose from Public Data Sets: Choose from public datasets like Jewellery Data set (Images), Gender Data Set (Images), Question or Sentence Data Set (Text), Numeric Data Set (CSV) or upload your data.

Type category(label) for the files (images/text file) that you have uploaded and click on submit to begin upload. Wait for some time till our web app uploads all the files. You can upload images for as many categories as possible.

Start Training - Push the button, to start the training. Now Metaverse’s intelligent backend will start with processing the data that you have uploaded and preparing it for the training.

2. In the sense of machine learning, explain the "No Free Lunch" theorem.

🡪The "No Free Lunch" theorem is a concept in machine learning and optimization theory that highlights the limitations of universal algorithms. The theorem states that, when averaged over all possible problems, no single machine learning algorithm performs better than any other.

The key idea behind the "No Free Lunch" theorem is based on two main assumptions:

1. Diverse problem landscapes: It assumes that all possible problem spaces are equally likely. In other words, problems vary greatly in their characteristics, structures, and complexities.

2. Uniformity of performance: When averaged over all possible problems, all algorithms perform equally well. This means that if one algorithm outperforms another on a particular problem, there must be other problems where the second algorithm performs better.

The theorem has profound implications for machine learning practitioners. It suggests that when choosing a machine learning algorithm for a specific problem, you cannot rely solely on its performance in other domains or on common beliefs about which algorithms are generally better. Instead, you should carefully analyse the problem's characteristics and dataset to select an algorithm that suits it best.

3. Describe the K-fold cross-validation mechanism in detail.

🡪K-fold cross-validation is a popular technique used to evaluate the performance of machine learning models and to assess their generalization ability on unseen data.Here's a step-by-step description of the K-fold cross-validation mechanism:

1. Data Splitting: The first step is to divide the original dataset into K subsets (or folds) of approximately equal size. Each fold represents a partition of the data, and it should have a similar distribution of classes and other relevant characteristics as the overall dataset.

2. Model Training and Testing: For each fold, the K-fold cross-validation process is performed as follows:

a. Select a Fold for Testing: One fold is used as a test set, and the remaining K-1 folds are used as the training set.

b. Model Training: The machine learning model is trained on the training set using the chosen algorithm and its hyperparameters.

c. Model Testing: The trained model is then evaluated on the test set to measure its performance. This evaluation produces metrics like accuracy, precision, recall, F1 score, etc., depending on the nature of the problem (classification, regression, etc.).

3. Iterative Process: Steps 2a to 2c are repeated K times, with each fold serving as the test set exactly once. As a result, each data point in the dataset is used for both training and testing, ensuring a comprehensive assessment of the model's performance.

4. Performance Metrics Aggregation: After performing K iterations, the performance metrics (e.g., accuracy) obtained in each iteration are averaged to provide a single evaluation score. This average score represents the overall performance of the model.

5. Model Selection and Hyperparameter Tuning: K-fold cross-validation can also be used to perform model selection and hyperparameter tuning. By trying different algorithms or hyperparameter combinations in each fold, you can identify the model setup that performs best on average across all folds.

4. Describe the bootstrap sampling method. What is the aim of it?

🡪The bootstrap sampling method is a resampling technique used in statistics and machine learning to estimate the variability and uncertainty associated with a dataset or statistical model. It involves creating multiple random samples (known as bootstrap samples) from the original dataset by drawing data points with replacement. The primary aim of the bootstrap method is to obtain robust estimates of statistical parameters and to assess the reliability of the model's predictions.

Here's a step-by-step description of the bootstrap sampling method:

1. Original Dataset: Start with a dataset containing N data points (observations).

2. Random Sampling with Replacement: To create a bootstrap sample, randomly select N data points from the original dataset, allowing for duplicates (sampling with replacement). As a result, some data points may appear multiple times in the bootstrap sample, while others may not be selected at all.

3. Sample Size: The size of the bootstrap sample is the same as the size of the original dataset (N). However, due to the random sampling with replacement, each bootstrap sample will be slightly different.

4. Parameter Estimation: Perform the analysis or computation of interest on each bootstrap sample. This could involve calculating statistical measures (e.g., mean, median, variance) or fitting a machine learning model to each sample.

5. Variability Estimation: The variability among the results obtained from different bootstrap samples reflects the uncertainty in the estimates. This variability can be used to calculate confidence intervals, standard errors, or other measures of uncertainty associated with the parameter of interest.

The main aim of the bootstrap sampling method is to approximate the underlying distribution of a statistical parameter or to obtain reliable estimates of model performance without making strong assumptions about the population distribution. It is particularly useful in situations where obtaining large amounts of additional data may be costly or impractical.

5. What is the significance of calculating the Kappa value for a classification model? Demonstrate how to measure the Kappa value of a classification model using a sample collection of results.

🡪Kappa value or Cohen's Kappa coefficient is an evaluation metric for classification models. Its significance as an evaluation metric is that it can be used to evaluate multi class classification models and also works on models trained on imbalanced datasets(scores like accuracy scores fail for imbalanced datasets).

In simpler words It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Cohen's kappa is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

6. Describe the model ensemble method. In machine learning, what part does it play?

🡪Ensemble methods or ensemble machine learning models are models where more than one models are being used spontaneously to produce better results than individually trained models. Ensemble modelling is a process where multiple diverse models are created to predict an outcome, either by using many different modelling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

The three main classes of ensemble learning methods are bagging, stacking, and boosting.

7. What is a descriptive model's main purpose? Give examples of real-world problems that descriptive models were used to solve.

🡪A descriptive model is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. As opposed to predictive models that predict a target of interest, in a descriptive model, no single feature is more important than any other. In fact, because there is no target to learn, the process of training a descriptive model is called unsupervised learning.

It is used in customer classification as real life problem .

8. Describe how to evaluate a linear regression model.

🡪Evaluation of a linear regression model can be done using R-square. R square is calculated as the sum of squared errors in predictions made, divided by summation of all sum of squares. R square measures how much of the change in target variable can be explained by the linear regressor. Its value ranges from 0 to 1 where 0 means poor performance and 1 means good. Some other techniques which can be used to evaluate a linear regression model are:

1.Mean squared error

2.Root mean squared error

3.Mean absolute error

4.Mean absolute percentage error

9. Distinguish :

1. Descriptive vs. predictive models

🡪Descriptive models are built to identify trends and underlying patterns.

Predictive models are built to predict a dependent variable value.

Most of descriptive models are built using unsupervised machine learning.

Most of predictive models are built using classification and regression models.

Example for descriptive model: Finding why consumers are engaging more with a social media post.

Example for predictive model: Predicting the chances of cancer in a patient.

2. Underfitting vs. overfitting the model

🡪Underfitting is a situation arising when the hypothesis is way too simple, or when the machine learning model is way too simple to produce good results.

Overfitting is a situation arising when the hypothesis is way too complex, or when the machine learning model is way too complex to produce good results.

Underfitting causes a model to produce poor results due to heavily simplified algorithm reacting lightly to changes in the unseen data for independent variables from the training data.

Overfitting makes a model produce poor results due to slightest variations in the unseen data for independent variables from the training data

Underfitting is also called High Bias.

Overfitting is also called High variance

3. Bootstrapping vs. cross-validation

🡪Bootstrap sampling is a method of sampling in which the repeated sampling is done with replacement using a data D in random draws over which machine learning models are trained for better performance.

Cross validation is a method used to check the efficacy of the machine learning model on test data.

End goal of bootstrapping is to reduce overfitting and increase performance.

End goal of cross validation is only to produce test scores to check efficacy of model

Bootstrapping is best employed in Random Forest Classifier.

Cross Validation is best employed using K-fold cross validation technique.

10. Make quick notes on:

1. LOOCV.

🡪Leave-One-Out Cross-Validation (LOOCV) is a special case of k-fold cross-validation where k is equal to the number of data points in the dataset. In LOOCV, for each iteration, a single data point is left out as the validation set, while the remaining data points are used as the training set to fit the model. This process is repeated for all data points in the dataset, and the model's performance is evaluated by averaging the results of all iterations.

2. F-measurement

🡪The F-measure, also known as the F1-score, is a popular performance metric used in binary classification tasks to assess the balance between precision and recall. It combines these two metrics into a single score, providing a more comprehensive evaluation of a classifier's performance, especially when dealing with imbalanced datasets.

3. The width of the silhouette

🡪 The silhouette width is a measure used to evaluate the quality of a clustering solution in unsupervised machine learning. It quantifies how well the data points within each cluster are separated from each other compared to the points in other clusters. A higher silhouette width indicates better-defined clusters, while a lower silhouette width suggests that the clusters are not well-separated or might be overlapping.

4. Receiver operating characteristic curve

🡪 The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model, particularly in machine learning and statistics. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds, providing insights into the model's ability to distinguish between the two classes (positive and negative).